Computer design of idealized β-motifs

Andrzej Kolinski[a)]
University of Warsaw, Department of Chemistry, Pasteura 1, 02-093 Warsaw, Poland and
The Scripps Research Institute, Department of Molecular Biology, La Jolla, California 92037

Wojciech Galazka
University of Warsaw, Department of Chemistry, Pasteura 1, 02-093 Warsaw, Poland

Jeffrey Skolnick
The Scripps Research Institute, Department of Molecular Biology, La Jolla California 92037

(Received 8 May 1995; accepted 11 September 1995)

A lattice model of protein conformation and dynamics is used to explore the requirements for the de novo folding from an arbitrary random coil state of idealized models of four and six-member β-barrels. A number of possible conjectures for the factors giving rise to the structural uniqueness of globular proteins are examined. These include the relative role of generic hydrophilic/hydrophobic amino acid patterns, the relative importance of the specific identity of the hydrophobic amino acids that form the core of the protein and the possible role played by polar groups in destabilizing alternative, misfolded conformations. These studies may also provide some insights into the relative importance of short range interactions, cooperative hydrogen bonding and tertiary interactions in determining the uniqueness of the native state, as well as the cooperativity of the folding process. Thus, these simulations may provide guidelines for the early stages of the protein design process. Possible applications to the general protein folding problem are also briefly discussed. © *1995 American Institute of Physics.*

I. INTRODUCTION

Lattice models of polypeptide conformation and dynamics are a useful tool for exploring aspects of the complex thermodynamics and folding kinetics of proteins. Indeed, a large variety of reduced or simplified models of proteins[1–23] have been studied.[24] Roughly speaking, these protein models can be divided into two classes. There are extremely schematic, simplified models that presumably reproduce some of the most basic features of proteins. These can be studied in great detail and provide the possibility of a rather rigorous test of some general theoretical questions concerning protein stability, cooperativity of interactions, and the nature of the folding process.[3,17–19] However, because of their highly schematic nature, they may miss important physical features of real proteins such as the role of hydrogen bonding in determining protein conformation and the relative entropy of compact states. On the other hand, there are intermediate resolution models that try to reproduce, with varying levels of accuracy, the geometrical and sequence details[14,20,23,25–28] with their ultimate objective being the folding of real protein sequences. It is expected that for these more accurate models, the energy landscape, dynamic features, and geometry of modeled structures more closely mimic the behavior of real proteins. Their disadvantage is that the numerical studies (due to larger computational costs) have to be less extensive. In this paper, we focus on the application of such models to the folding of idealized sequences of β-proteins.

Recently, we proposed an intermediate resolution model based on the use of a fine lattice approximation to the conformational space of polypeptide chains.[25–28] The resulting Cα representation of the polypeptide main chain is geometrically rather accurate. All three-dimensional protein structures from the Brookhaven Protein Data Bank (PDB) (Refs. 29 and 30) can be fitted to the lattice with an error in the range of 0.6–0.7 Å root mean square deviation (rms) (Ref. 23) for the Cα atoms. The side chain rotamers were represented as a single sphere, located at the center of mass of the side chain in its particular isomeric state. The geometric accuracy of the side chain representation is on the level of 1–2 Å rms. The hydrogen bonds were implicitly defined using a variant of the Levitt–Greer method[31] of secondary structure assignment. The force field of the model was based on several potentials of mean force derived from the statistics of interactions seen in the PDB structures. The resulting approximations in the geometrical representation and interaction scheme lead to a circa 2–3 Å rms accuracy for the representation of the entire protein backbone. This means that two low energy structures, with essentially the same side chain contact map[12] and the same pattern of model hydrogen bonds may differ by 2–3 Å rms. Starting from sequence information alone and a randomly selected initial denatured state, a low to moderate resolution folded state of several small proteins has been obtained.[14,25–28] Generally, the method was more successful for helical proteins. β-motifs can only be reproducibly obtained in the case of very short sequences and very simple topology, as was demonstrated in the folding simulations of crambin.[26] For more complex β-type structures, the model could not distinguish between various topologies and/or it was unable to find the lowest energy, ''nativelike'' state in a reasonable amount of computer time. Part of the reason for this failure may be due to the increased topological complexity of β-proteins and part may be due to the unphysically large flexibility of β-strands exhibited by these models.

It is important to understand the reasons for these limi-

[a)]Address for correspondence: The Scripps Research Institute, Department of Molecular Biology, 10666 North Torrey Pines Road, MB1, La Jolla, California 92037.

 0021-9606/95/103(23)/10286/12/$6.00

tations of the model, which was quite successful for simple helical proteins and small (also helical) macromolecular assemblies.[27] Therefore, the short range interaction scheme was re-examined.[32] The most important update was a more straightforward treatment of the peptide bond atoms. As was recently demonstrated by Oldfield and Hubbard,[33] the Cα trace based definition of secondary structure is no less specific than the frequently used phi−psi Ramachandran map.[34] Consequently, the positions of peptide bond atoms between the ith and $i+1$st alpha carbons are precisely defined by two consecutive planar angles of the Cα trace and one dihedral angle. The short range angular correlations between peptide bond plates are typical of those found in regular secondary structure; thus, the correlations of the Cα vectors and correlations of the side chain directions provide a better representation of secondary structure propensities encoded in the amino acid sequence. Additionally, there is a simple and fast way to define geometrical criteria for main chain hydrogen bonding. The resulting improved model provides a geometric representation that more closely mimics the geometry found in the native state of globular proteins. In particular, by including the positions of the amide hydrogens and carbonyl oxygens in the calculation of the hydrogen bond energy, the registration of both parallel and antiparallel β-sheets is substantially improved.

Understanding the requirements for the formation of simple structural motifs is a necessary step along the way towards the de novo computer modeling of more complex proteins. Studying such simplified sequences in the context of intermediate resolution protein models may also provide more general insights into the thermodynamics and dynamics of protein folding. Among the questions addressed are the following: Can a general hydrophobic/hydrophilic pattern comprised of two kinds of amino acids yield unique native like states? There are some suggestions that variation in identity of the hydrophobic amino acids in the core of the protein is a prerequisite for a unique native state. What is the role of surface hydrophilic residues and charged groups? Are they necessary to help eliminate alternative topologies? What is the role of turns? Can they influence the outcome for the stability of the final topology? By exploring the effect of such sequence variations on the uniqueness of the final folded conformation, the objective of this paper is to provide some insights into the factors responsible for formation of a unique topology in globular proteins. However, it is very likely that the sequences we consider do not have unique side chain packing and that the low temperature state corresponds to the molten globule state of proteins.[35–37] The experimental design of sequences which exhibit side chain fixation has also proven to be difficult;[35] thus, these studies do not focus on the problem of side chain fixation. Rather, we examine questions associated with the formation of a dense collapsed state with proper secondary structure, and with a unique topology of connections of these secondary structural elements.

II. DESCRIPTION OF THE MODEL

The geometrical details of the fine lattice reduced representation of polypeptide chains were recently described.[32] Here, we briefly outline the model for the reader's convenience. The reference state for the definition of backbone atoms and each single ball side chain is provided by the Cα backbone that is confined to the lattice. The n-residue polypeptide is represented by $n+1$ vectors $a\cdot\mathbf{v}$ that connect n consecutive Cαs, and two terminal caps. These vectors belong to the set of 90 vectors $\{\mathbf{v}\}=\{[3,1,1],...[3,1,0],...[3,0,0],...[2,2,1],...[2,2,0],.\}$; $a=1.22$ Å and is obtained from the best fit of lattice Cαs to PDB structures. For each amino acid, we define a side chain rotamer library. The number of model rotamers of residue i depends on its identity and on backbone local geometry as defined by $\mathbf{v}_{i-1}$ and $\mathbf{v}_i$. The centers of mass of the model rotamers coincide (by construction) with an accuracy of 1.0 Å with respect to real side group rotamers. On the basis of the frequency of various rotamers in protein structures, a mean field statistical potential can be defined. In addition, there are terms that reflect the intrinsic tendency of amino acids to adopt a given kind of secondary structure. Some of these terms are specific to the particular sequence under consideration; others are generic and are designed to adjust the lattice model so that it adopts a set of proteinlike conformational states. Hydrogen bonding is also included. While the free energy of peptide backbone hydrogen bond formation is in reality probably comparable to the free energy of backbone hydrogen bond formation with water, what is most certainly true is that the absence of any backbone hydrogen bonds is very costly. Thus, in both the models and in real proteins, this is an essential term that serves to eliminate many distorted, nonproteinlike backbone conformations and is a very important geometric regularizing term. Then, there are terms that reflect the tendency of individual amino acids to be buried or exposed to water. Finally, there are pair interactions that reflect the preference of amino acid pairs to interact with each other. All statistical potentials described below are available by anonymous ftp.[42]

A. Short range interactions

The short range interactions that have been designed to reproduce intrinsic secondary structure propensities have been recently described.[32] The sequence dependent local conformational propensities are characterized by the following potentials of mean force:

$$E_s=\sum \epsilon(A_i,A_{i+1},r^{2*}_{i-1,i+2}), \tag{1}$$

with

$$r^{2*}=\mathrm{sign}[(\mathbf{v}_{i-1}\otimes\mathbf{v}_i)\cdot\mathbf{v}_{i+1}]r^2_{i-1,i+2},$$

$$r^2_{i-1,i+2}=(\mathbf{v}_{i-1}+\mathbf{v}_i+\mathbf{v}_{i+1})^2,$$

where Σ denotes the summation along the chain, A_i is residue type at position i, and $\mathbf{v}_i$ is the virtual bond vector from the ith to $i+1$st Cα. The local polypeptide conformation is defined by three consecutive backbone vectors; $\mathbf{v}_{i-1}$, $\mathbf{v}_i$, and $\mathbf{v}_{i+1}$. $r^{2*}_{i-1,i+2}$ is the ''chiral'' square of distance between the corresponding chain vertices. ''Chiral'' means a negative sign for left handed conformations and a positive sign for right handed conformations, respectively.

We also include an amino acid pair specific potential that reflects the angular correlations between side chain vectors[25]

$$E_{\text{sg-local}}=\sum \epsilon_k[A_i,A_{i+k},\cos(\Theta_{i,i+k})] \quad k=1,2,3,4, \tag{2}$$

where $\Theta_{i,j}$ is the angle between the side group vectors (the vector from the Cα to the center-of-mass of the current rotamer) of residues i and j. The potential has the form of a histogram with an angular interval equal to 36 deg (and a range of 0–180 deg).

There are also three generic (i.e., amino acid independent) terms describing the short range interactions. They correct the distribution conformational states of the lattice chain, thereby enforcing a ''proteinlike'' distribution of states. The first such term is of the form

$$E_g=\sum \epsilon_g(\mathbf{v}_{i-1},\mathbf{v}_i,\mathbf{v}_{i+1}), \tag{3}$$

where E_g is defined in the same spirit as the short range sequence specific contribution, E_s. Here, the exact number of occurrences of a particular triplet of vectors in a lattice fitting to a set of native protein structures is used. It is stored as a projection into six bins of the chiral end-to-end distance for a particular conformation of the chain fragments. The zero of energy corresponds to the average frequency of vector triplets seen in the database.

Additionally, there are corrections to the distribution of the end-to-end distance of four vector fragments. The distribution should be peaked at small distances corresponding to helical (and turnlike) states and at large distances corresponding to expanded (β-type) states. This stands in contrast to the athermal lattice chains whose distribution has a maximum at intermediate distances. The correction is of the following simple form:

$$E_\eta=\sum \eta_i(r^2_{i-2,i+2}), \tag{4}$$

where

$$\eta_i=-1 \quad \text{for } r^2_{i-2,i+2}<35,$$

$$\eta_i=-1 \quad \text{for } r^2_{i-2,i+2}>75,$$

$$\eta_i=0 \quad \text{otherwise.}$$

The ''proteinlike'' stiffness of the main chain is simulated by enforcing strong orientational correlations of the peptide bond plates[32] (only *trans* conformations of the peptide bond are assumed),

$$E_p=\sum [\cos(\mathbf{h_i},\mathbf{h_{i+2}})+\cos(\mathbf{h_i},\mathbf{h_{i+4}})], \tag{5}$$

where $\cos(\mathbf{h_i},\mathbf{h_j})$ denotes the cosine between the ith and jth vectors defining the orientation of the peptide bond plates (the vectors between the amide hydrogen and the carbonyl oxygen). We note that the positions of the backbone atoms are well defined by the local geometry of the Cα chain.

The total short range energy of the model chain is computed as

$$E_{\text{local}}=E_s+E_{\text{sg-local}}+E_g+E_\eta+E_p. \tag{6}$$

Due to the presence of long range interactions, the sequence specific secondary structure preferences are suppressed in comparison to the scaling employed in the previous study of the model that only possessed short range interactions.[32] Thus, all the energetic terms are equally weighted.

B. Model of hydrogen bonds

Having the positions of the oxygen and hydrogen atoms of the peptide bonds defined and stored in a local coordinate system provided by three consecutive Cα backbone vectors, it is possible to define model hydrogen bonds as a simplified Coulombic interaction between these atoms. The hydrogen and oxygen atoms interact via the following potential:

$$\epsilon_{\text{H-bond}}=q_{\text{H}}(1-f_{\text{H}})/[r_{\text{O,H}}+2\cdot\exp(-r^2_{\text{O,H}})], \tag{7}$$

where f_{H}, the angular factor, is of the following form:

$$f_{\text{H}}=[0.77-\cos(\mathbf{r}_{\text{O}i,\text{H}j},\mathbf{r}_{\text{O}i,\text{H}i})]^2 + [0.77-\cos(\mathbf{r}_{\text{O}i,\text{H}j},\mathbf{r}_{\text{O}j\text{H}j})]^2. \tag{8}$$

q_{H} is an arbitrary scaling factor for the strength of the hydrogen bond in model peptides that implicitly accommodates partial charges, local dielectric constant, etc. $\mathbf{r}_{\text{O}i,\text{H}j}$ is the vector between the oxygen in peptide plate j and the hydrogen in peptide plate i. The above scheme reproduces the average geometry of H-bonds in proteins with reasonable accuracy. Due to the lack of a hydrogen atom in the peptide bond, proline residues can participate in one main chain hydrogen bond. This model definition of the H-bonds, when applied to the original (off-lattice) Cα, traces of PDB structures recovers almost all main chain H-bonds as assigned by the Kabsch–Sander[38] method. However, a substantial number of local hydrogen bonds are omitted.

Parallel and antiparallel β-structures have a different pattern of hydrogen bonds when the identity of individual hydrogen bonds (e.g., residue 6 has a hydrogen bond with residue 44) is associated with the amino acids. If, however, one associates the numbering of amide hydrogens and carbonyl oxygens with consecutive peptide bonds rather than with the original amino acid units, then the hydrogen bond pattern becomes the same for both types of β-structures. This relabeling has an additional advantage. The resulting relabeled hydrogen bond pattern is qualitatively the same for helical and all types of β motifs. Consequently, the cooperativity of the hydrogen bond network can be simply introduced. We add a cooperative contribution for all consecutive pairs of hydrogen bonds, i.e., when there is an i,j hydrogen bond and an $i+1,j+1$ hydrogen bond (also an $i-1,j-1$, or an $i-1,j+1$, or an $i+1,j-1$ bond) at the same time. This requires the definition of a threshold value for the formation of a hydrogen bond, which is set to $0.20q_{\text{H}}$ in the context of Eqs. (7) and (8). (This number was obtained in order to make this hydrogen bond definition compatible with that given by Kabsch and Sander.) The contribution per cooperative H-bond is assumed to be equal to the above threshold value. For hydrogen bonds participating in regular structural motifs (β-sheets or α-helices), the cooperative contribution is

TABLE I. Numerical values of pairwise interaction energy parameters.

	BCK	GLY	ALA	SER	CYS	VAL	THR	ILE	PRO	MET	ASP	ASN	LEU	LYS	GLU	GLN	ARG	HIS	PHE	TYR	TRP
BCK	−1.0	−0.1	−0.2	0.1	−0.5	0.3	0.4	0.5	0.7	0.4	1.4	0.8	0.7	1.7	2.3	1.1	1.5	1.2	1.1	1.1	0.8
GLY	−0.1	0.4	0.1	0.1	0.3	0.3	0.1	0.5	0.7	0.4	0.0	0.3	0.5	1.6	1.2	0.5	0.9	0.8	0.1	−0.5	−1.1
ALA	−0.2	0.1	−0.4	0.1	−0.5	−0.7	0.1	−0.8	0.4	−0.7	0.5	0.3	−0.7	1.7	1.5	0.4	1.2	−0.2	−1.2	−0.9	−0.9
SER	0.1	0.1	0.1	−0.6	0.2	0.5	−0.2	0.7	0.6	0.8	−2.0	−0.3	0.7	1.4	−0.4	0.6	0.8	−0.7	0.3	0.1	0.4
CYS	−0.5	0.3	−0.5	0.2	−12.3	−0.6	0.0	−0.7	−0.5	−1.5	1.1	0.3	−0.7	2.1	1.9	0.7	1.2	−1.1	−2.8	−1.0	−0.6
VAL	0.3	0.3	−0.7	0.5	−0.6	−1.3	0.0	−1.4	0.5	−1.0	2.4	1.3	−1.2	2.0	2.0	0.8	1.3	0.6	−2.6	−0.9	−2.6
THR	0.4	0.1	0.1	−0.2	0.0	0.0	−0.5	0.0	0.6	0.3	−1.1	−0.5	0.4	1.9	0.2	0.2	0.8	−0.9	0.1	−0.2	0.4
ILE	0.5	0.5	−0.8	0.7	−0.7	−1.4	0.0	−1.6	0.6	−1.5	2.2	1.8	−1.4	2.0	1.9	1.0	1.3	0.7	−2.5	−1.6	−3.0
PRO	0.7	0.7	0.4	0.6	−0.5	0.5	0.6	0.6	0.8	0.2	2.0	1.1	0.8	3.2	2.5	0.5	1.6	−1.1	−0.4	−2.8	−3.3
MET	0.4	0.4	−0.7	0.8	−1.5	−1.0	0.3	−1.5	0.2	−2.7	2.1	1.1	−1.3	1.8	1.4	0.9	1.6	−1.5	−4.4	−2.7	−4.5
ASP	1.4	0.0	0.5	−2.0	1.1	2.4	−1.1	2.2	2.0	2.1	1.2	−1.1	2.6	−2.7	2.5	0.4	−5.9	−2.0	2.4	−1.1	0.4
ASN	0.8	0.3	0.3	−0.3	0.3	1.3	−0.5	1.8	1.1	1.1	−1.1	−1.8	1.8	1.2	0.5	−0.8	0.7	0.1	0.9	−0.7	−1.5
LEU	0.7	0.5	−0.7	0.7	−0.7	−1.2	0.4	−1.4	0.8	−1.3	2.6	1.8	−1.5	2.4	2.2	0.6	1.5	−0.1	−2.5	−1.1	−3.1
LYS	1.7	1.6	1.7	1.4	2.1	2.0	1.9	2.0	3.2	1.8	−2.7	1.2	2.4	5.2	−3.0	2.0	5.8	3.6	2.6	−1.2	−0.6
GLU	2.3	1.2	1.5	−0.4	1.9	2.0	0.2	1.9	2.5	1.4	2.5	0.5	2.2	−3.0	4.1	2.0	−6.4	−1.5	2.9	−0.2	0.9
GLN	1.1	0.5	0.4	0.6	0.7	0.8	0.2	1.0	0.5	0.9	0.4	−0.8	0.6	2.0	2.0	1.2	−0.3	0.9	−0.1	−1.5	−2.0
ARG	1.5	0.9	1.2	0.8	1.2	1.3	0.8	1.3	1.6	1.6	−5.9	0.7	1.5	5.8	−6.4	−0.3	0.2	0.0	0.3	−2.5	−7.0
HIS	1.2	0.8	−0.2	−0.7	−1.1	0.6	−0.9	0.7	−1.1	−1.5	−2.0	0.1	−0.1	3.6	−1.5	0.9	0.0	−1.4	−3.5	−5.3	−5.9
PHE	1.1	0.1	−1.2	0.3	−2.8	−2.6	0.1	−2.5	−0.4	−4.4	2.4	0.9	−2.5	2.6	2.9	−0.1	0.3	−3.5	−5.7	−2.9	−4.6
TYR	1.1	−0.5	−0.9	0.1	−1.0	−0.9	−0.2	−1.6	−2.8	−2.7	−1.1	−0.7	−1.1	−1.2	−0.2	−1.5	−2.5	−5.3	−2.9	−2.9	−2.8
TRP	0.8	−1.1	−0.9	0.4	−0.6	−2.6	0.4	−3.0	−3.3	−4.5	0.4	−1.5	−3.1	−0.6	0.9	−2.0	−7.0	−5.9	−4.6	−2.8	−6.4

doubled and serves to propagate regular secondary structure. The physical meaning of this semiempirical approach is very similar to the propagation parameter of the Zimm–Bragg statistical mechanical model for the helix–coil transition,[39] or the similar Mattice[40] model of the β-sheet to random coil transition. Here, both helical and β-sheet conformational transitions are treated identically.

C. Long range interactions

Long range tertiary interactions contain two contributions; both are based on the statistics of interactions seen in high resolution protein structures from the PDB. The first contribution is a one body, centrosymmetric potential that reflects the tendency of hydrophobic residues to be buried inside the globule and the tendency of hydrophilic residues to be exposed to solvent. The solvent molecules implicitly contribute to the potential of mean force,[41]

$$E_1=\sum \epsilon_1[r(A_i)/S_0], \tag{9}$$

with

$$S_0=2.2n^{0.38} \ (\text{in Å}), \tag{10}$$

where S_0 is the expected radius of gyration of a single domain protein consisting of n amino acids in its native form. $r(A_i)$ is the distance of the center of mass of the ith side group from the center-of-mass of the entire chain. The potential is derived from the statistics of single domain proteins and is used in the form of a histogram. The reference state is a randomly packed globule of average composition. The values of ϵ_1 are available via anonymous ftp.[42]

All pairwise tertiary interactions are neglected up to the fourth neighbors down the chain, since these short range interactions are already accounted for by the hydrogen bond potential and secondary structure preferences. The pair interactions beyond the fourth neighbor are derived from the statistics of the database. For residues i and j,

$$E_{ij}=\begin{cases} E^{\text{rep}}, & \text{for } r_{ij}<R_{ij}^{\text{rep}} \\ \epsilon_{ij}, & \text{for } R_{ij}^{\text{rep}}<r_{ij}<R_{ij}, \text{ and } \epsilon_{ij}>0 \\ f\epsilon_{ij}, & \text{for } R_{ij}^{\text{rep}}<r_{ij}<R_{ij}, \text{ and } \epsilon_{ij}<0 \end{cases} \tag{11}$$

$$f=1.0-[\cos^2(\mathbf{u_i},\mathbf{u_j})-\cos^2(20°)]^2, \tag{12}$$

where R_{ij}^{rep} and R_{ij} are the cutoff values for strong excluded volume interactions and for square-well, soft pairwise interactions, respectively. R_{ij}^{rep} equals the average contact distance $\overline{R_{ij}}$ minus two standard deviations of this value, whereas R_{ij} is equal to $\overline{R_{ij}}$ plus three standard deviations. The average pairwise contact distances and the standard deviations are available via anonymous ftp.[42] E^{rep} is the penalty for hard sphere overlap and is equal to 4–5 kT. ϵ_{ij} is the pairwise interaction energy (for all the combinations of the side groups and the backbone units). The procedure for the derivation of ϵ_{ij} is found in the Appendix, and the parameters themselves are found in Table I. For side groups, it is moderated by an angular factor that reflects the average preferred packing angle between interacting secondary structure elements. For nonstructured fragments (as detected during the simulation), $f=1$; while for structured fragments (helices or β-strands), f is given by Eq. (12), where $\mathbf{u}_i=\mathbf{r}_{i+2}-\mathbf{r}_{i-2}$, and $\mathbf{r}_k$ are the coordinates of the kth Cα.

D. Scaling of the contributions to the potential

The force field described in the preceding sections contains several potentials of mean force that must be scaled. This scaling is, to some extent, arbitrary, but would produce nonphysical behavior outside of a range of values of the parameters. The values of these scaling factors were chosen

after analysis of simulations of several sequences of real proteins. The objective was to obtain as accurate as possible secondary structure in the collapsed state and as small as possible rms for correctly folded fragments. In addition, in the unfolded state, hydrophobic clusters could not be too stable, and only a marginal amount of regular secondary structure (5–10 %) could be present. If hydrophobic clusters are too stable, then the molecule becomes kinetically trapped and will not fold. Conversely, if they are too weak, then the molecule never becomes compact. In particular, the one body term tends to generate a reasonable distribution of hydrophobic and hydrophilic residues with respect to the interior of the compact protein. If pairwise interactions are too weak, then one sees association of charges having identical sign both on the surface and in the interior of the molecule. If secondary structure terms are turned off, then collapse to random compact states having little secondary structure is observed. This is in contrast to the hypothesis of Dill and co-workers[17] that compaction will induce secondary structure. If intrinsic secondary preferences are too strong, then too much secondary structure will be present in unfolded state. Empirically, we found that an approximately 1:1 distribution of the energy between the short and long range contributions, respectively, can produce denatured states with marginal secondary structure and yield compact conformations with regular secondary structure and a reasonable distribution of hydrophobic and hydrophilic residues. For pairwise interactions, the scaling factor is 0.325; for the H-bond energy, it is equal to 2.5. Otherwise, all scaling factors are equal to 1. A more detailed analysis of the effects of various contributions to the potential on the nature of the folding transition and the character of the resulting compact conformations will be discussed in greater detail in a future publication.

E. Simulation algorithm

The simulation algorithm used in this work is very similar to that described previously[25] except for the energy updates described above. A standard asymmetric Metropolis scheme[43] is used which employs a set of local micromodifications of the chain conformation and small distance motions of larger parts of the model chain. All simulations started from an expanded random coil state, which is different for each run. During the simulation, the temperature is gradually lowered from a temperature of 1.7 (well above the folding temperature), to $T=1.0$ ($T=0.8$ for the four member β-barrel, see below) that is below the folding temperature. The folding temperature could be determined from the energy (and energy fluctuation) changes, as well as from the conformational characteristics (e.g., chain collapse). Unlike researchers performing exhaustive enumeration studies of relative short chains on very simple lattices, we cannot be sure that the lowest free energy (or energy) state is obtained. Thus, a series of simulations is run, and the reproducibility of the structures between independent simulations is checked. We assign native states based on the idea that they exhibit the lowest minimum and average energy during the course of the simulation and that the putative native conformation is reproducibly folded.

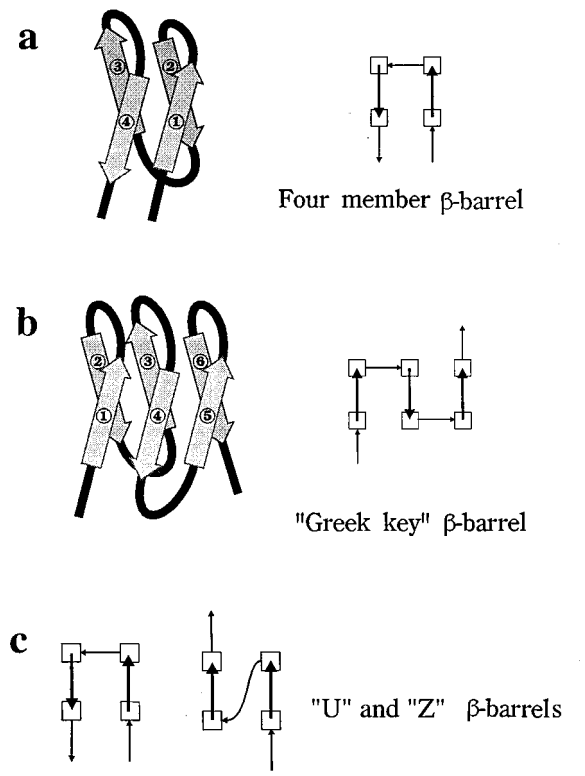

FIG. 1. Schematic representation of the two designed β-type folding motifs (a and b). The corresponding amino acid sequences are given in Tables II and IV. In the topology diagrams shown on the right-hand side, the thick (thin) arrows indicate the connections on the top (bottom) of barrels. The left-hand side depicts the corresponding ribbon type topological diagrams. Two possible topologies of the four member β-barrel are also shown (c).

III. RESULTS AND DISCUSSION

Figures 1(a) and 1(b) show schematic drawings (ribbon models and topology diagrams) of the two β motifs that are the target of the sequence designs studied in this work. Even though four member β-barrels are not seen in nature, they form the core of the Greek key β-barrels. Consequently, a four member β-barrel was chosen as the first example for sequence design. In particular, in these simulations, we focused on the ''U'' topology as opposed to the ''Z'' topology, shown in Fig. 1(c). We then proceeded to build on insights obtained from this simple case to design a sequence which folds to the more complicated Greek key topology, for which domain 1 of chymotrypsin serves as the prototype.

A. Design of a minimal four member β-barrel

The first very simple sequence we designed is shown in the second column of Table II, along with the expected secondary structure which is shown in column five. The basic design idea was to choose an alternating pattern of identical hydrophobic, VAL, and hydrophilic residues, SER, to form the putative beta strand regions. Both VAL and SER are expected to be beta forming residues. We also chose three identical sets of turns sequences, formed by GLY-PRO-ARG, since these residues are more hydrophilic. Eleven independent simulations were performed. Of these, seven adopted very irregular, low temperature conformations, four adopted a single sheet structure and one formed a Z-type four mem-

TABLE II. Sequences of designed four member β-barrels.

Residue #	β4a	β4b	β4c	Secondary structure type
1	HIS	GLY	GLY	coil
2	SER	SER	GLU	β
3	VAL	VAL	VAL	β
4	SER	SER	SER	β
5	VAL	VAL	PHE	β
6	SER	GLU	GLU	β
7	VAL	VAL	VAL	β
8	GLY	GLY	GLY	turn
9	PRO	GLY	GLY	turn
10	ARG	VAL	VAL	turn
11	SER	LYS	LYS	β
12	VAL	VAL	PHE	β
13	SER	SER	SER	β
14	VAL	VAL	VAL	β
15	SER	GLY	LYS	β
16	GLY	GLY	GLY	turn
17	PRO	GLY	GLY	turn
18	ARG	GLY	GLY	turn
19	VAL	VAL	VAL	β
20	SER	SER	SER	β
21	VAL	PHE	PHE	β
22	SER	SER	SER	β
23	VAL	VAL	VAL	β
24	GLY	GLY	GLY	turn
25	PRO	GLY	GLY	turn
26	ARG	VAL	VAL	turn
27	SER	SER	SER	β
28	VAL	PHE	PHE	β
29	SER	SER	SER	β
30	VAL	VAL	VAL	β
31	SER	SER	SER	β
32	GLY	GLY	GLY	coil

ber barrel, which was the lowest energy structure observed by 10 kT. The topology of a Z barrel is schematically depicted in Fig. 1(c). The results of simulations on the first sequence (β4a) seem to indicate that there are several isoenergetic (or nearly isoenergetic) structures. This situation is somewhat similar to the two isoenergetic structures observed in experimentally designed four helix bundles.[44,45] This lack of reproducibility may be due to a number of competing effects. Hydrogen bonding tends to drive the system to a single sheet structure, whereas hydrophobic interactions tend to generate collapsed structures. The uniform pattern of hydrophilic and hydrophobic residues has multiple conformations where the strands can pack with identical pair interaction energy. In other words, this sequence lacks interactions that destabilize alternative folded conformations. Furthermore, the hydrophobic core may be too small to generate unique stable structures. Finally, because of the large number of unburied hydrophobic residues, a single sheet structure would not be expected to be stable in nature; this points out the need to improve the burial potential. Work in this direction is now in progress.

Could we select a single topology of the barrel by proper modification of the sequence? The redesigned sequence, β4b (see Table II, column 3), has more flexible turns (GLY-residues). We introduced two oppositely charged residues (LYS and GLU) in the first two putative strands that are designed to form a part of the hydrophilic surface of the first sheet. The hydrophobic core of the model protein is better defined by introducing two PHE residues in the middle of the second putative sheet, on its hydrophobic face. This should not only enhance the hydrophobicity of the model, but the inclusion of nonidentical hydrophobic residues should break the degeneracy of the compact conformations.

In a series of five runs, the behavior of this redesigned sequence was found to be qualitatively different from sequence β4a. In two cases, the expected topology of the barrels has been obtained. These final structures consist of two well defined sheets (strands 1 and 2 form the first β-sheet and strands 3 and 4 form the second β-sheet) which exhibit well defined packing of the hydrophobic residues. A ''Z'' type four member barrel is obtained in one run, and single sheet structures are obtained in the other two simulations. Clearly, the introduced sequence mutations exhibit a large influence on the behavior of the model system. However, the misfolded states once again reflect the interplay of H-bonding plus associated cooperativity, which prefer a single sheet arrangement and the burial/hydrophobic interactions which favor a two sheet structure. This competition can also be noticed from the comparison of various contributions to the total energy. For example, one of the single sheet structures has the lowest H-bond energy because it has a larger network of model H-bonds.

Due to apparently insufficient hydrophobic interactions of sequence β4b to drive formation of the hydrophobic core, we examined a third sequence β4c (shown in the fourth column of Table II). β4c contains two additional PHE residues, and one additional pair of oppositely charged residues in the putative N-terminal sheet. As indicated by Table III, the β4c sequence behaves more like a real protein. Four of the ten simulations produce ''U'' type four member barrels; a representative example is shown in Fig. 2(a). Three of the ten simulations produce ''Z'' type four member barrels; a representative example is shown in Fig. 2(b). One of these Z-shaped barrels is a mirror image of the other two. The remaining simulations yield distorted beta barrels, the lowest energy of which is a ''U''-shaped barrel, whose sheets are at roughly right angles to each other, see Fig. 2(c).

When the three lowest energy structures are considered (see Table III), this sequence rapidly folds to a well-defined ''U'' shaped four member barrel, with a higher degree of reproducibility. This is not only evident on the level of global topology, but also on the level of side chain contact maps (which reflect better defined packing of the ''protein's'' hydrophobic core) and the small rms between structures (in the range of 3 Å). Based on the side chain contact pattern, the fourth ''U'' shaped four member barrel did not develop a full hydrophobic core, and therefore, its energy was somewhat higher. Thus, because of its highest reproducibility and lowest energy, the U-shaped four member β-barrel is tentatively assigned to be the proper low energy state for the β4c sequence. However, based on these results, the topology is not uniquely defined.

TABLE III. Energy of the final state from β4c simulations.

Simulation	Final E	Rotamer	Local backbone	H bond energy	Long range	Topology[a]
1	−174.1	−13.0	−79.7	−31.2	−50.1	U
2	−176.8	−9.0	−79.5	−35.1	−53.0	U
3	−147.3	−8.3	−72.0	−28.1	−38.8	Z
4	−155.1	−10.4	−71.1	−30.5	−42.9	Z
5	−163.2	−10.1	−78.4	−27.4	−47.2	⋯
6	−169.6	−13.8	−80.1	−31.0	−44.5	U
7	−145.4	−12.0	−71.0	−28.4	−34.3	⋯
8	−158.3	−12.3	−73.6	−33.1	−39.1	U
9	−169.2	−13.6	−77.4	−37.1	−40.9	Z
10	−162.5	−9.7	−81.4	−30.3	−40.9	⋯

[a]Z, z-shaped connections; U, the u-shaped connections.

B. Simulations of six-member Greek key β-barrels

The experiments on the minimal four-member beta barrel show that substantial sequence specificity is required in order to obtain an almost uniquely folded state. However, the failure to design a unique topology is partially caused by the inability to form a well defined and completely buried hydrophobic core in such a small structure. Consequently, we examined a sequence which is expected to adopt a six-member β-barrel fold. In particular, the sequence designed to adopt the Greek key topology depicted in Fig. 1(b) is shown in column two of Table IV and contains 45 residues. It incorporates the design insights obtained from our study of sequence β4c. The putative turns are comprised of GLY linkers chosen because of their conformational flexibility. The first putative sheet consists of strands 1, 4, and 5 and the second putative sheet consists of strands 2, 3, and 6. The hydrophobic core is anchored by two pairs of PHE residues in central strands 3 and 4. The hydrophilic surfaces are built from charged residues whose location is designed to disfavor alternative topologies.

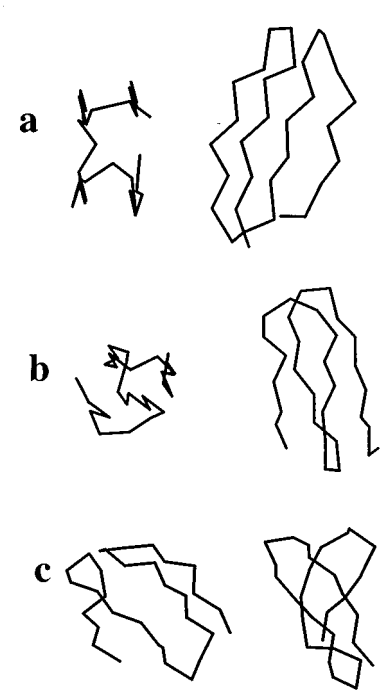

FIG. 2. Representative final conformations in two different views for the β4c sequence given in Table II. (a) The lowest energy U-shaped barrel (top and side view). (b) The Z-shaped barrel (top and side view). (c) Misfolded structure composed of two almost orthogonal minimal β-sheets.

The designed sequence was then subject to 14 independent folding simulations, each starting from a different random expanded state and driven by slow temperature annealing from $T=1.5$ to $T=1.1$. The resulting energies including the contribution of the various components are shown in Table V. In all the simulations, when the temperature is below 1.4, the proper topology formed (in some runs, the proper fold formed and dissolved several times). In 11 out of 14 runs, at $T=1.1$, the final state was locked in the expected fold. In two cases, the final state has topological errors and is clearly misfolded (simulations #6 and #11). These folds have the noticeably higher energy by about 30–55 kT. In both cases the errors occurred near the C-terminus of the polypeptide chain. In one case, the mirror image topology has been obtained (simulation #7). Again, it has higher energy than the proper topology. The properly folded states (11 of 14) from various runs are the same within the resolution of the model. It is noteworthy that the energy of correct folds from various simulations ranges from $-263kT$ to $-235kT$. Moreover, the individual contributions to the final energy can vary quite a bit for these states (compare, for example, simulation #1 with simulation #13; both have almost the same total energy, but the long range contributions differ by $10kT$). The contribution that most clearly differentiates folded from misfolded states is the hydrogen bond energy. These observations strongly indicate that while the correctly folded structures are unique on the level of topology (almost the same pattern of hydrogen bonds) the details of the model side chain packing and local main chain conformations are less accurate.

A typical Cα trace for the final state at a temperature $T=1.1$ is shown in Fig. 3. The rms deviations between the pairs of independently folded structures are on the level 1.5–3.5 Å. This is very close to the level of resolution of the present model. In all cases, the hydrophobic core is well defined and unique (all the PHE residues are in contact with the other PHE residues; however, when larger rms deviations between structures are present, some fluctuations are apparent), and the pattern of hydrogen bonds is very reproducible (more than 80% similarity between the final correctly folded

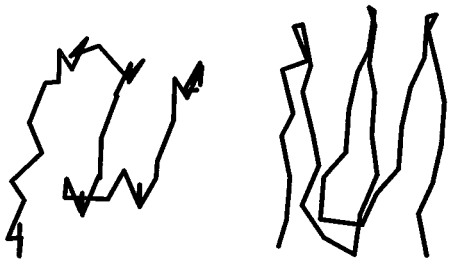

FIG. 3. The snapshot of the representative final state from the folding simulation of the sequence designed to adopt a Greek-key topology (top and side view).

structures). The same level of reproducibility is seen for contact maps[14] of the side groups.

The behavior of the model system during the slow simulated annealing is very interesting. In Fig. 4, we display a typical folding trajectory. Folding in this case is seen to occur by an on site construction mechanism[46] (also referred to as a hydrophobic zipper model[47]), where secondary structure formation is accompanied by the formation of tertiary contacts. In the particular case discussed here, the system starts from a relatively expanded random coil state. The entire trajectory contains 200 snapshots taken every time unit. Here, we display 8 snapshots showing typical folding events. One time unit corresponds to 2500 $*m$ attempted micromodifications per residue, with m equal to a small constant (which is about 10) that reflects various types of local motions[25] attempted in a single MC cycle. We remind that the particular MC moves[25,32] involve a different number of chain units, and a single cycle of the algorithm attempts various moves in a random order. At high temperatures at the beginning of simulated annealing, the motion of the model chain is very fast. The first three snapshots (at $t=10$, 25, and 27) show completely uncorrelated conformations. However, even at this

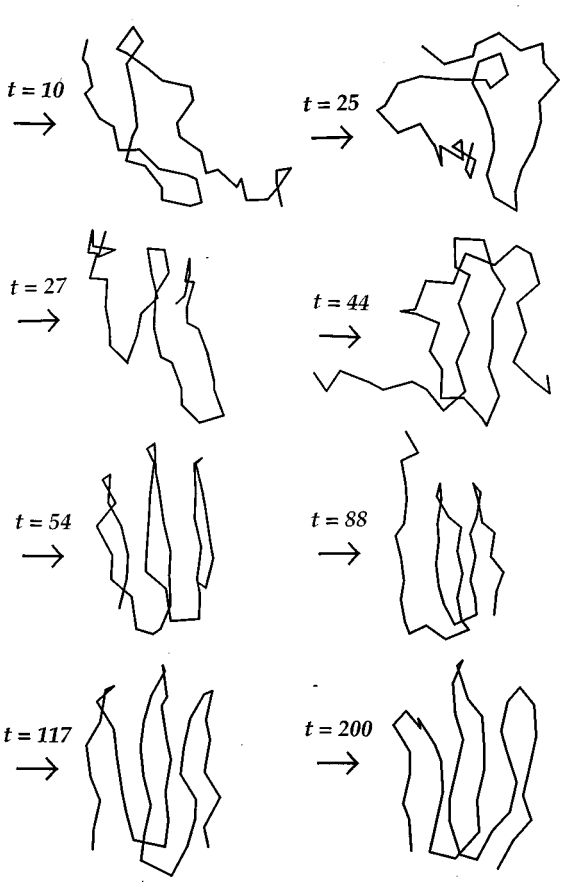

FIG. 4. The schematic representation of a representative folding trajectory of the Greek-key barrel. The numbers indicate simulation time of the particular snapshots. The time unit corresponds to $2500*m$ (m is a small integer, see the description of the Monte Carlo algorithm) attempted conformational jumps per residue.

TABLE IV. Sequence of designed six-member β barrel.

Residue #	3	Type
1	GLY	coil
2	VAL	β
3	ASP	β
4	VAL	β
5	ASP	β
6	VAL	β
7	GLY	coil
8	GLY	coil
9	GLY	coil
10	VAL	β
11	ASP	β
12	VAL	β
13	ASP	β
14	VAL	β
15	GLY	coil
16	GLY	coil
17	PHE	β
18	ARG	β
19	PHE	β
20	ARG	β
21	VAL	β
22	GLY	coil
23	GLY	coil
24	GLY	coil
25	VAL	β
26	ARG	β
27	PHE	β
28	ARG	β
29	PHE	β
30	GLY	coil
31	GLY	coil
32	VAL	β
33	ASP	β
34	VAL	β
35	ASP	β
36	VAL	β
37	GLY	coil
38	GLY	coil
39	GLY	coil
40	VAL	β
41	ASP	β
42	VAL	β
43	ASP	β
44	VAL	β
45	THR	coil

stage, sometimes there are short lived elements of secondary structure; as, for example, the β-hairpin seen at $t=27$. The hairpin consists of putative strands 4 and 5 of the expected Greek key fold. Subsequently, the model polypeptide chain undergoes a very rapid collapse to a more dense state (snapshot at $t=44$). The first well-defined elements of secondary structure are formed in the middle part of the chain and involve strands 2 and 3. However, in other trajectories strands 4 and 5 associate first. At the same time, the system forms a nucleus of a hydrophobic core containing strongly hydrophobic PHE residues. The end strands (usually 1 and 6, and sometimes also strand #5) are the last to assemble. These intermediates usually form and dissolve several times. In this particular trajectory, the correct fold appears for the first time at $t=54$. It then dissolves a couple of times and again adopts all or the most of the proper fold, with the chain ends (especially the C-terminus) being the last to assemble. After $t=100$ (see snapshots at $t=117$, and $t=200$, that marks the end of the simulation), the folded structure undergoes mostly minor conformational fluctuations. However, in some trajectories, an almost complete unfolding/folding process could be observed even at $t>150$. Here, the final structure ($t=200$) has a well defined β-type pattern of hydrogen bonds that defines two sheets (strands 1, 4, and 6 in the first sheet and strands 2, 3, and 6 in the second sheet).

Additional insight into the nature of the folding process (another simulation) is provided by Fig. 5, where we plot the rms deviation of the Cα trace (after the best superposition) from the final state (at the time equal to 200) vs simulation time. The system starts from a random, expanded conformation and very rapidly undergoes collapse to denser states. These are characterized by a rms from native which is in the range of 6 Å. It is of interest to note that different structures with a rms in the range of 6 Å from the ''native state'' usually have qualitatively different folds. In contrast, when the rms deviation is about 3 Å, the mutual arrangements of the strands are native like; there are only minor deviations in the hydrogen bond network and the packing of the side chains.

The simulation depicted in Fig. 5 shows that there are a few transitions into and out of the native conformation. In this run, the native structure (albeit with substantial local fluctuations) lasts from $t=50$ to $t=70$; then, upon a larger

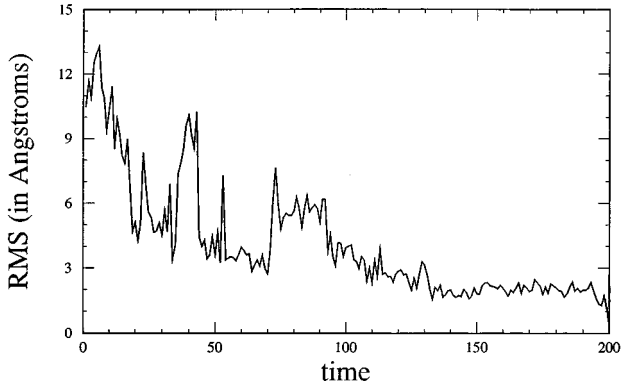

FIG. 5. Plot of the distance Cα trace rms from the final state of the representative folding experiment on the designed Greek-key barrel sequences as a function of simulation time. The time unit corresponds to $2500*m$ (m small integer, see the text) attempted conformational jumps per residue. The initial temperature is $T=1.5$, the final temperature is equal to 1.1, and changes in linear fashion with time.

thermal fluctuation, it dissolves. At lower temperatures, after another collapse to the native fold (roughly near $t=100$ in this run), the system spends quite a long time adopting a better packing arrangement of the side chains and readjusting the hydrogen bond network. The plateau near 2 Å rms indicates that there are no additional substantial conformational changes and that a unique low temperature state (within the resolution of the model) has been adopted. Comparison with simulations on four member barrels, indicates that the folding in the six-member barrel is more cooperative.

The above statement that independent simulations converge to the same structure is further substantiated by comparison of two independent trajectories. In Fig. 6, we plot the rms deviation between corresponding snapshots from two independent simulations as a function of MC time. At high temperatures (at the beginning of simulated annealing), the trajectories are completely uncorrelated, with an rms in the range of 12 Å between corresponding conformations. A local minimum around $t=80$ to $t=90$, reflects the coincidental time (which is random) of two folding attempts. After $t=150$, both trajectories show the native state, and the rms

TABLE V. Energy for the final states from the Greek key simulations.

Simulation	Final E	Rotamer	Local backbone	H bond energy	Long range	Topology
1	−258.9	−19.6	−106.3	−51.9	−81.1	correct
2	−241.5	−22.0	−105.6	−42.4	−71.4	correct
3	−235.1	−20.4	−110.3	−46.2	−58.3	correct
4	−241.8	−16.6	−111.4	−45.8	−68.0	correct
5	−247.0	−21.4	−107.6	−47.4	−70.7	correct
6	−218.8	−23.1	−96.4	−37.9	−61.5	misfolded
7	−240.5	−16.8	−98.6	−44.8	−80.4	image
8	−249.3	−23.6	−119.5	−45.7	−60.4	correct
9	−252.7	−23.2	−104.4	−47.7	−77.4	correct
10	−257.2	−21.2	−112.6	−49.8	−73.5	correct
11	−208.1	−19.3	−91.1	−34.3	−63.4	misfolded
12	−250.9	−23.3	−113.8	−50.9	−62.9	correct
13	−256.1	−22.0	−112.3	−50.9	−71.0	correct
14	−263.1	−21.5	−115.7	−51.2	−74.7	correct

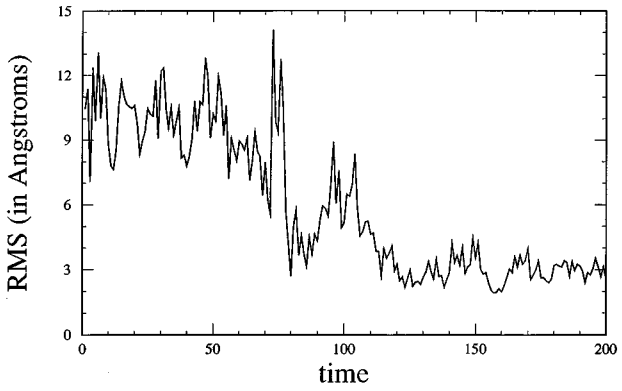

FIG. 6. Representative plot of distance rms between corresponding snapshots of two folding trajectories for the Greek-key sequence as a function of time (see the text).

distance between two snapshots taken at identical simulation times decays slowly to values in the range of 2–3 Å.

These results show that our design of Greek key barrel is reasonably successful. The model polypeptide cooperatively folds with high reproducibility to a low temperature state which is structurally unique in the context of this model. Of course, there is no guarantee that in reality this sequence is foldable, nor is the accuracy of the present model sufficient to differentiate native conformations from molten globule states. Nevertheless, these simulations demonstrate that model beta proteins having a close resemblance to real structures are obtained when the sequence strongly encodes for secondary structure, has hydrophilic residues that eliminate alternative folds and has hydrophobic residue diversity that reduces the degeneracy of side chain packing in the protein core.

IV. SUMMARY AND CONCLUSION

In the present paper, we have attempted to design sequences that adopt simplified beta motifs. While we realize that the conclusions obtained are dependent on how accurately the model mimics the features of real proteins, with this caveat in mind, the following qualitative features for beta protein design criteria emerge. Much as in the case of designed helical proteins studied both experimentally and in an earlier version of this model, structural uniqueness requires that the hydrophobic core contain a number of different hydrophobic residue types. This is required to insure that a unique registration of the side chains is achieved. Here, this is accomplished by using both VAL and PHE as the putative core residues. Interestingly, as has been suggested by Harbury *et al.*,[48] we also find that interactions between hydrophilic residues are important in that they destabilize misfolded structures. In our case, they are located on the outside of the protein and act to favor the correct Greek key topology and disfavor alternative strand packing arrangements. Finally, we have found that glycines can be used to mark turn regions. This would indicate that under suitable conditions, the native topology can be determined by the packing interactions between secondary structural elements such as beta strands. Such a result has also been seen experimentally in the redesign of some naturally occurring proteins.[49]

A final important conclusion from the present study is that the force field proposed for the reduced representation of the polypeptides reflects some general features of the complex interactions that stabilize the native conformation in globular proteins. The model responds in expected ways to ''mutations'' of amino acid sequence. The information encoded in the sequence is at least qualitatively transmitted into a potential that properly selects not only the structural class of the folded state, but also finer topological and packing details. The low temperature states are ''proteinlike'' on the level defined by the resolution of the proposed model. The best example is the designed six-member, Greek key β-barrel. The optimized sequence folds with high reproducibility to a unique (within the 2–3 Å resolution of the model) ''nativelike'' state. Thus, for a highly idealized β-sequence, successful de novo folding from arbitrary random conformations to a structure that has much in common with the native state of real proteins has been demonstrated.

While these results are encouraging, it is important to emphasize that the finer atomic details are absent in the present reduced model. Consequently, it is rather difficult to conclude whether the proposed sequences would, in reality, be able to adopt unique hydrophobic core packing, with side chain fixation, that is characteristic of native proteins.[36] It is possible that they can only achieve a ''molten globule'' state.[36,37] The problem of the kinetic accessibility of the ''nativelike'' state needs further investigation. Another potential problem in real experiments may arise from protein aggregation; an effect which is entirely ignored here. Nevertheless, it appears that the crude engineering of proteins could be aided by the proposed method.

In future work, the model will be used to design more complex structural motifs and to predict of the structure and folding pathways of additional small natural globular proteins.

ACKNOWLEDGMENTS

This work was supported in part by NIH Grant No. GM-37408, by NIH Fogarty International Collaboration Grant No. TW-00418, and by the University of Warsaw Grant BST 472-34/94. A. Kolinski is an International Research Scholar of the Howard Hughes Medical Institute. We thank the anonymous referee for suggestions that helped us to improve the presentation of this work.

APPENDIX

The present version of the pair potential is based on the idea that the effective interaction between pairs of residues should reflect the average interaction energy between the heavy atoms that constitute the interacting partners. Thus, a protein is divided into twenty three classes of interacting residue types$\{j\}$. The 23 classes consist of the backbone heavy atoms, the 20 naturally occurring amino acid side chains, crosslinked cystine and ligated histidine. Within each class, there are a number of members, m_j, comprising the heavy atoms. In the backbone, for example, there are four

members, γ, consisting of the alpha carbon, the C′ carbon, the carbonyl oxygen, and the amide nitrogen. For a given side chain type, the members are treated distinctly, so that the interaction of a beta carbon belonging to valine may have a different pair potential from that of a beta carbon in phenylalanine. This was done so as to include the possibility of specific geometric effects (e.g., the fact that a beta carbon in alanine will likely have more interresidue interactions than a beta carbon in tyrosine, which may be partially shielded due to geometric factors).

The pair potential between heavy atoms γ and δ belonging to side chains i and j, respectively, is estimated from

$$\epsilon(\gamma\delta,ij) = -\ln[N_{\text{observed}}(\gamma\delta,ij)/N_{\text{expected}}(\gamma\delta,ij)]. \tag{A1}$$

Here, $N_{\text{observed}}(\gamma\delta,ij)$ is the observed number of contacts in a database of 228 PDB globular protein structures, and $N_{\text{expected}}(\gamma\delta,ij)$ is the expected number of contacts, if the groups of heavy atoms randomly mix with no preferential pair interactions whatsoever. In a folded protein environment, which includes secondary structure and a collection of groups of different size and shape, the calculation of $N_{\text{expected}}(\gamma\delta,ij)$ is extremely difficult.

To obtain an estimate of $N_{\text{expected}}(\gamma\delta,ij)$, we follow in the spirit of a previous generalization of Flory–Huggins theory,[25] but we now explicitly focus on the determination of a potential of mean force between heavy atom pairs. Each heavy atom is assumed to have a total coordination number of sites $z=7.8$. This number is near the maximum number of contacts (corrected for very infrequent occurrences) seen in the PDE database and is obtained by assuming that a pair of heavy atoms is in contact if their distance is less than or equal to 4.5 Å. Each chemical bond that forms reduces the number of the possible interaction sites by one. Thus, the backbone carbonyl oxygen has $z-1$ sites, a backbone carbonyl carbon has $z-3$ sites, and an alanine methyl group has $z-1$ sites. Let $z_{\gamma,j}$ be the number of available interaction sites of group γ in residue class j.

In the database of L protein structures, let N_j be the total number of examples of class type j. Then, the total number of possible contacts is

$$N = \sum_{j=1}^{23} \sum_{\gamma=1}^{m_j} N_j z_{\gamma,j}. \tag{A2}$$

Note that N will in general be greater than or equal to the total number of observed contacts, N_{obs}, since there are surface residues as well as interior residues whose coordination sphere is not completely saturated.

We now introduce the contact fraction of group γ in residue class j, $\phi_{\gamma j}$,

$$\phi_{\gamma j} = N_j z_{\gamma,j} \Big/ \sum_{k=1}^{23} \sum_{\gamma=1}^{m_k} N_k z_{\gamma,k}. \tag{A3}$$

If all the heavy atoms have the same coordination number, then the contact fraction is equal to the mole fraction. In general, this is not the case, and the contact fraction acts to account for different coordination numbers of the groups present in proteins.

For group γ in residue class i, the total number of possible interactions is $N_i z_{\gamma i}$. In the Bragg–Williams approximation, the probability that each site interacts with group type δ of class j $(\neq i)$ is $\phi_{\delta j}$. Thus, the expected number of $\gamma\delta,ij$ contacts,

$$N_{\text{expected}}(\gamma\delta,ij) = N_i z_{\gamma i}\phi_{\delta j} = N\phi_{\gamma i}\phi_{\delta j}. \tag{A4}$$

Similarly, the number of contacts between identical groups is

$$N_{\text{expected}}(\gamma\gamma,ij) = N\phi_{\gamma i}^2/2. \tag{A5}$$

The factor of 2 corrects for overcounting.

Using Eq. (A4) or (A5) in Eq. (A1), provides for the interaction energy between heavy atom pairs $\gamma\delta$ of classes i and j, $\epsilon(\gamma\delta,ij)$. In the lattice folding simulations, side chains are represented as single spheres located at the side chain centers of mass; thus, we require the average pair interaction energy between residues (or classes) of types i and j, $\epsilon(i,j)$. To obtain this quantity, we average $\epsilon(\gamma\delta,ij)$ over all heavy atom contacts in the original database. Specifically, let the lth protein in the database $(l=1,L)$, contain nres(l) residues. Let A_k denote the class of residue type $(=2-23)$ at position k in the sequence $[k=1,\text{nres}(l)]$. All backbone groups are treated identically for convenience; however, in general this need not be true. $k>(<)0$ denotes the side chain (backbone) located at position k in the sequence. We define a generalized contact matrix between all pairs of heavy atoms $\gamma\delta$ of classes A_i and A_j, at positions i and j in the sequence

$$C(\gamma\delta,A_iA_j) = \begin{cases} 0 & \text{if heavy atoms } \gamma\delta, \quad A_iA_j \text{ are not in contact} \\ 1 & \text{if heavy atoms } \gamma\delta, \quad A_iA_j \text{ are in contact} \end{cases}. \tag{A6}$$

Similarly, we define the side chain-side contact function X by

$$X(A_iA_j) = \min\left[\sum_{\gamma=1}^{m_{A_i}} \sum_{\delta=1}^{m_{A_j}} C(\gamma\delta,A_iA_j), 1\right]. \tag{A7}$$

That is, X equals 1 if there is at least one heavy atom–heavy atom contact, and it is zero, otherwise. Thus, the average pair energy $E(i,j)$ is obtained from

$$\epsilon(i,j) = \frac{\sum_{l=1}^{L}\sum_{i=-\text{nres}(l)}^{\text{nres}(l)}\sum_{j=-\text{nres}(l)}^{\text{nres}(l)}\sum_{\gamma=1}^{m_{A_i}}\sum_{\delta=1}^{m_{A_j}} C(\gamma\delta ij)\epsilon(\gamma\delta ij)}{\sum_{l=1}^{L}\sum_{i=\text{nres}(i)}^{\text{nres}(i)}\sum_{j=-\text{nres}(j)}^{\text{nres}(j)} X(A_iA_j)}. \tag{A8}$$

The numerator is simply the total interaction energy between pairs of heavy atom groups in the database, and the denominator is simply the total number of contacting pairs of classes of residue types i and j which have at least one heavy atom pair in contact.

[1] M. Go and H. A. Scheraga, Biopolymers **23**, 1961 (1984).
[2] M. Levitt, Curr. Opinion Struct. Biol. **1**, 224 (1991).
[3] K. A. Dill, Curr. Opinion Struct. Biol. **3**, 99 (1993).
[4] M. Karplus and E. Shakhnovich, *Protein Folding* (Freeman, New York, 1992), pp. 127–195.
[5] C. Wilson and S. Doniach, Proteins **6**, 193 (1989).
[6] M. Levitt and A. Warshel, Nature **253**, 694 (1975).
[7] A. T. Hagler and B. Honig, Proc. Natl. Acad. Sci. USA **75**, 554 (1978).
[8] I. D. Kuntz, G. M. Crippen, P. A. Kollman, and D. Kimelman, J. Mol. Biol. **106**, 983 (1976).
[9] J. Skolnick and A. Kolinski, J. Mol. Biol. **221**, 499 (1991).
[10] J. Skolnick and A. Kolinski, Science **250**, 1121 (1990).
[11] A. Godzik, A. Kolinski, and J. Skolnick, J. Comput. Aided Mol. Design **7**, 397 (1993).
[12] A. Godzik, A. Kolinski, and J. Skolnick, J. Mol. Biol. **227**, 227 (1992).
[13] D. G. Covell, Proteins **14**, 409 (1992).
[14] A. Kolinski, A. Godzik, and J. Skolnick, J. Chem. Phys. **98**, 7420 (1993).
[15] J. Skolnick, A. Kolinski, and R. Yaris, Proc. Natl. Acad. Sci. USA **86**, 1229 (1989).
[16] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[17] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993); Macromolecules **22**, 4529 (1989).
[18] E. I. Shakhnovich and A. M. Gutin, Proc. Nat Acad. Sci. USA **90**, 7195 (1993).
[19] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[20] M-H. Hao and H. A. Scheraga, J. Phys. Chem. **98**, 4940, 9882 (1994).
[21] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[22] A. Godzik, J. Skolnick, and A. Kolinski, Proc. Natl. Acad. Sci. USA **89**, 2629 (1992).
[23] A. Godzik, A. Kolinski, and J. Skolnick, J. Comput. Chem. **14**, 1194 (1993).
[24] T. E. Creighton, *Proteins, Structure and Molecular Properties* (Freeman, San Francisco, 1984).
[25] A. Kolinski and J. Skolnick, Proteins **18**, 338 (1994).
[26] A. Kolinski and J. Skolnick, Proteins **18**, 353 (1994).
[27] M. Vieth, A. Kolinski, C. L. Brooks III, and J. Skolnick, J. Mol. Biol. **237**, 361 (1994).
[28] J. Skolnick, A. Kolinski, C. L. Brooks III, A. Godzik, and A. Rey, Current Biol. **3**, 414 (1993).
[29] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Simanouchi, and M. Tasumi, J. Mol. Biol. **112**, 535 (1977).
[30] PDB Q. Newsletter **63**, (1993).
[31] M. Levitt and J. Greer, J. Mol. Biol. **114**, 181 (1977).
[32] A. Kolinski, M. Milik, J. Rycombel, and J. Skolnick, J. Chem. Phys. **103**, 4312 (1995).
[33] T. J. Oldfield and R. E. Hubbard, Proteins **18**, 324 (1994).
[34] G. N. Ramachandran and V. Sasisekharan, Adv. Protein Chem. **23**, 283 (1968).
[35] D. P. Raleigh and W. F. DeGrado, J. Am. Chem. Soc. **1992**, 10 079.
[36] K. Kuwajima, Proteins **6**, 87 (1989).
[37] O. B. Ptitsyn, R. H. Pain, G. V. Semisotnov, E. Zerovnik, and O. I. Razgulyaev, FEBS **262**, 20 (1990).
[38] W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
[39] B. H. Zimm and J. R. Bragg, J. Chem. Phys. **31**, 526 (1959).
[40] W. L. Mattice, Annu. Rev. Biophys. Biophys. Chem. **18**, 93 (1989).
[41] T. L. Hill, *An Introduction to Statistical Thermodynamics* (Dover, New York, 1960).
[42] A. Kolinski and J. Skolnick, *Parameters of statistical potential*. Available by ftp from public directory, scripps.edu (pub/MCSP), 1995.
[43] *Monte Carlo Methods in Statistical Physics*, edited by K. Binder (Springer, Berlin, 1986).
[44] T. M. Handel and W. F. DeGrado, Biophys. J. **61**, A265 (1992).
[45] T. M. Handel, S. A. Williams, and W. F. DeGrado, Science **261**, 879 (1993).
[46] J. Skolnick and A. Kolinski, Annu. Rev. Phys. Chem. **40**, 207 (1989).
[47] A. A. Rashin, Stud. Biophys. (Berlin) **77**, 177 (1979).
[48] P. B. Harbury, T. Zhang, P. S. Kim, and T. Alber, Science **262**, 1401 (1993).
[49] L. C. Wu, R. Grandori, and J. Carey, Protein Sci. **3**, 369 (1994).